Modeling and Simulation 建模与仿真, 2023, 12(4), 3770-3780
Published Online July 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.124345

Hans汉斯

基于双流卷积多注意力模型的行人意图识别研究

张晓斐，王孝兰*

上海工程技术大学机械与汽车工程学院，上海

收稿日期：2023年5月4日；录用日期：2023年7月11日；发布日期：2023年7月18日

摘　要

识别行人等弱势道路使用者的行为意图是自动驾驶汽车做出有效决策和控制动作保护行人和驾驶者安全的前提。本文设计了一种基于双流结构融合时空特征的行人过街意图识别模型(Dual-stream Convolutional Multi-Attention Model, DCMAM)。基于MobileNet引入空间注意力设计空间流卷积模块；基于膨胀3D卷积网络(Inflated 3D ConvNet, I3D)引入时空和空洞卷积设计时间流卷积模块；基于门控循环单元(Gate Recurrent Unit, GRU)搭建双向GRU网络，捕获时空交互信息；引入注意力机制设计双流融合模块。在数据集JAAD和PIE上的实验证明了模型的有效性，意图识别准确率相较于现有方法提高了7%。集成意图识别模型和硬件平台设计行人意图识别系统，通过实车实验验证了意图识别系统的稳定性和准确性。

关键词

行人意图识别，自动驾驶汽车，注意力，双流卷积神经网络，时空卷积

Research on Pedestrian Intention Recognition Based on Dual-Stream Convolutional Multi-Attention Model

Xiaofei Zhang, Xiaolan Wang*

School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai

Received: May 4th, 2023; accepted: Jul. 11th, 2023; published: Jul. 18th, 2023

Abstract

For autonomous vehicles to effectively make decisions and control actions to ensure the safety of pe-

*通讯作者。

文章引用: 张晓斐, 王孝兰. 基于双流卷积多注意力模型的行人意图识别研究[J]. 建模与仿真, 2023, 12(4): 3770-3780.
DOI: 10.12677/mos.2023.124345

destrians and drivers, they must be able to recognize the behavioral intention of vulnerable road users, such as pedestrians. This paper designs a pedestrian crossing intention recognizemodel (Dual-stream Convolutional Multi-Attention Model) based on the fusion of spatiotemporal features of the dual-stream network structure. Introducing spatial attention based on MobileNet to create the spatial flow convolution module; designing the time Stream convolution module by adding spatio-temporal convolution and atrous convolution based on Inflated 3D ConvNet (I3D); a bidirectional GRU network is constructed based on Gate Recurrent Unit (GRU) to capture spatio-temporal interaction information. The attention mechanism is introduced to design the dual-stream fusion module. Comparative experiments on the datasets JAAD and PIE demonstrate the effectiveness of the proposed method, with a 7% improvement in intention recognition accuracy compared to existing methods. Based on a hardware platform integrating an intention recognition network model, a pedestrian intention recognition system is created. The stability and accuracy of the intention recognition system is verified through real vehicle experiments.

Keywords

Pedestrian Intention Recognition, Autonomous Vehicles, Attention, Two-Stream Convolutional Neural Network, Spatio-Temporal Convolution

1. 引言

自动驾驶汽车要在愈发复杂的城市道路环境中实际部署，拥有实时预测行人、自行车等弱势道路使用者(Vulnerable Road User, VRU)的意图，并采取必要的行动保护其安全的能力是至关重要的[1]。

胡远志[2]将行人骨骼和方向分别处理，充分利用骨骼的 2D 信息，建立了双流自适应图卷积网络进行行人意图识别。杨彪[3]基于行人动作和环境条件等多种信息建立行人意图识别网络。曹昊天[4]基于长短时记忆网络设计了一种多特征融合的行人意图识别和轨迹预测方法。Chen [5]和 Lorenzo [6]引入行人边界框和人体骨架姿态作为行人过街意图预测的新特征源。SF-GRU [7]基于循环神经网络(Recurrent Neural Networks, RNN)搭建堆叠 RNN 模型分层融合五个特征源用于行人过街意图预测。PCPA [8]将四个特征源直接拼接再利用全连接层实现预测。随着特征源的增加，网络结构愈加复杂。PCIP [9]拥有五个特征源，还要设计复杂的融合架构保证网络的精度。然而 PCIP 只考虑隐含的时间特征，而光流天然包含的时间信息可以帮助识别和预测行人的过街意图。

为了解决以上问题，我们提出了一种基于双流网络结构融合时空特征的行人过街意图识别模型。基于图像序列获取光流、行人外观表征和全局上下文分别作为时空特征源。基于 MobileNet 结合空间注意力设计了空间流特征提取模块，基于 I3D 引入时空卷积和空洞卷积设计了时间流特征提取模块，搭建了基于注意力机制的双流融合模块。在数据集 JAAD 和 PIE 上与其他方法进行比较来证明本文提出模型的有效性。设计行人过街意图识别系统，以真实的城市道路条件为实验环境，验证意图识别系统的稳定性和准确性。

2. 双流卷积特征提取

由于大多数城市交通场景和行人的不可预测性，为了建立行人运动特征模型，在特征提取时对每个行人的观察数据进行采样。每个目标行人 i 都有一个固定长度 m 的连续图像序列，在本文中图像序列长度定为 16 帧。目标行人 i 基于给定连续图像序列提取的特征信息定义如下：

行人外观表征：基于数据集中标注的目标行人边界框放大固定倍数逐帧裁剪并重新调整大小得到行人外观表征，即：

$$R_{li} = \left\{ r_{li}^{t-m}, r_{li}^{t-m+1}, \cdots, r_{li}^{t} \right\} \tag{1}$$

其中目标行人 i 的边界框在 $t-m$ 帧图像中的坐标位置为：

$$l_i^{t-m} = \left\{ x_{it}^{t-m}, y_{it}^{t-m}, x_{ib}^{t-m}, y_{ib}^{t-m} \right\} \tag{2}$$

其中 $x_{it}^{t-m}, y_{it}^{t-m}$ 表示边界框左上角坐标，$x_{ib}^{t-m}, y_{ib}^{t-m}$ 表示边界框右下角坐标。

全局上下文：采用在 Cityscapes 数据集[10]上进行预训练的 DeepLab V3 [11]模型来提取语义掩码作为全局上下文，即

$$C_g = \left\{ c_g^{t-m}, c_g^{t-m+1}, \cdots, c_g^{t} \right\} \tag{3}$$

光流：利用 Flownet2 [12]模型预生成数据集中所有视频的光流信息，即

$$O = \left\{ o^{t-m}, o^{t-m+1}, \cdots, o^{t} \right\} \tag{4}$$

2.1. 空间流特征提取

为了提取图像信息中的空间特征，基于 MobileNet [13]和空间注意力[14]搭建空间流特征提取模块。如图 1 所示，它由标准卷积块，深度可分离卷积块和空间注意力融合块组成。预处理得到的全局上下文 C_g 和行人外观表征 R_{li} 经过卷积块处理后送入 Lambda 层融合，然后经过空间注意力层对特征图的空间信息进行权重再分配，随后经过平均池化层下采样得到空间流特征编码 S_C。

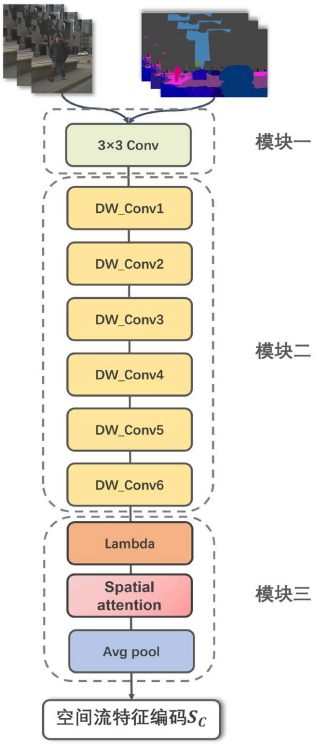


Figure 1. Spatial stream feature extraction module
图 1. 空间流特征提取模块

MobileNet 将标准卷积替换为深度可分离卷积，增加了批量归一化(Batch Normalization, BN)层和激活函数 Relu6。深度可分离卷积首先采用深度卷积对不同输入通道分别进行卷积，然后采用逐点卷积将前一步的输出结合，卷积效果类似于标准卷积，可以降低模型计算复杂度和模型大小，更适合自动驾驶汽车等算力受限的应用场景。

空间注意力用于特征图内不同位置特征信息的权重再分配，保持空间维度不变，压缩通道维度获得注意力权重重新分配网络关注度，使网络更关注目标位置。对注意力模块的输入从通道维度分别进行平均池化和最大池化，得到两个特征图并拼接，通过卷积层将其变为 1 通道的特征图，再经过 sigmoid 函数得到空间注意力的特征权重，最后将输出结果与原特征图相乘恢复输入尺寸。

2.2. 时间流特征提取

为了获取光流中的时间特征信息，基于 I3D [15]重新设计了时间流特征提取模块。如图 2 所示，它由十二个卷积块、四个最大池化层和一个全局平均池化层组成。后处理块由 Flatten 层和 Dense 层组成。

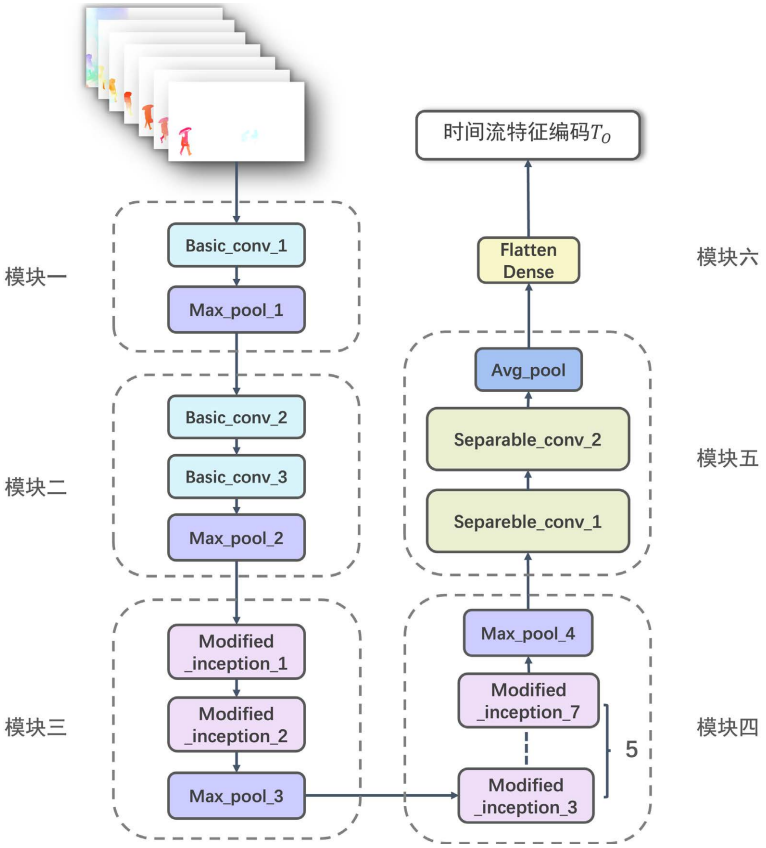


Figure 2. Temporal feature extraction module
图 2. 时间流特征提取模块

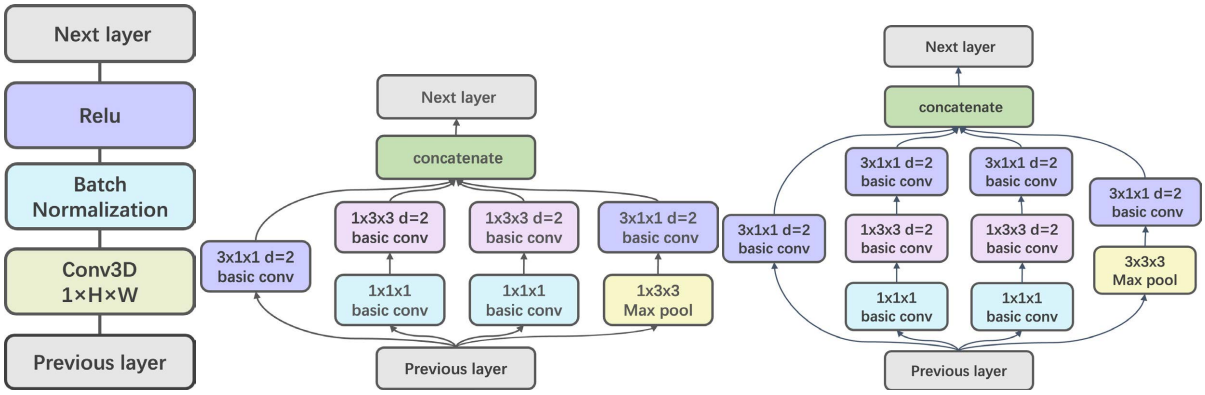
光流O送入网络经过卷积处理后，将特征向量输入后处理块得到时间流特征编码T_O。

卷积块共有三种类型，分别是基础卷积块、改进 inception 卷积块和时空卷积块，所有卷积层都添加 BN 层和非线性激活函数 Relu。改进 inception 和时空卷积块引入空洞卷积，空洞率设置为 2。

基础卷积块只对空间维度卷积，基础卷积块 1 保持步长为 2，只对空间特征进行下采样并提取底层较丰富的空间特征信息。卷积块结构如图 3(a)所示。

改进 inception 卷积块采用 2D 卷积核，并引入空洞卷积。模块三的最大池化层对时空特征信息进行下采样，融合并向后传播。卷积块结构如图 3(b)所示。

时空卷积块同时引入空洞卷积和分离卷积，将原 I3D 的 3D 卷积改为两个连续的 2D 卷积，首先学习空间特征，然后学习时间特征，将时间信息添加到所有分支，提高获取高层语义特征的效率，利用 1 × 1 × 1 的卷积层实现特征通道的降维减参，最后再将四个不同尺度的特征拼接然后输出特征向量，卷积块结构如图 3(c)所示。



(a) 基础卷积块　　(b) 改进 inception 卷积块　　(c) 时空卷积块

Figure 3. Convolutional block structure
图 3. 卷积块结构

3. 双流卷积多注意力模型设计

双流卷积多注意力模型结构如图 4 所示，它由空间流卷积模块，时间流卷积模块和双流融合模块组成。

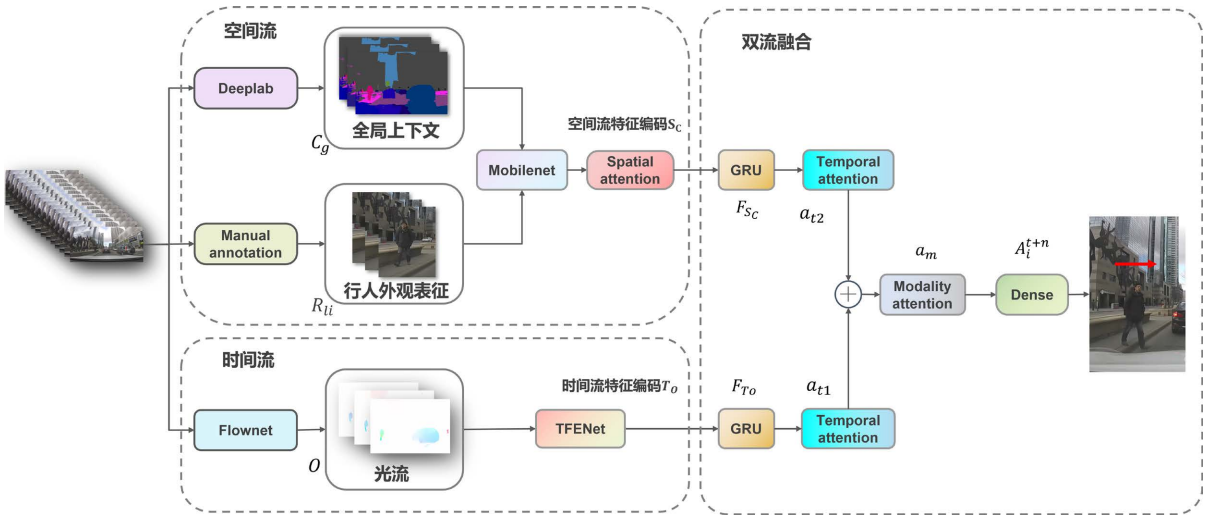


Figure 4. Two-stream convolutional multi-attention model
图 4. 双流卷积多注意力模型

在双流融合模块中引入注意力机制[16]，时间注意力用于序列特征内不同时序特征信息的权重再分配，提高神经网络对时序信息的利用率。注意力模块的输入为 GRU 输出的隐藏状态 $h_i = \{h_1, h_2, \cdots, h_e\}$。注意

力权重为：

$$\alpha_t(s)=align\left(h_t,\bar{h}_s\right)=\frac{\exp\left(score\left(h_t,\bar{h}_s\right)\right)}{\sum_{s'}\exp\left(score\left(h_t,\bar{h}_s\right)\right)} \tag{5}$$

其中基于特征内容的“得分”函数为 $score\left(h_t,\bar{h}_s\right)=h_t^{\mathrm{T}}\bar{h}_s$，$t$ 是序列长度，h_t 是解码器当前隐藏层状态，$\bar{h}_s$ 是编码器所有的隐藏层状态。注意力模块的输出向量为：

$$a_t=f\left(c_t,h_t\right)=\tanh\left(W_c\left[c_t,h_t\right]\right) \tag{6}$$

其中 W_c 是可训练权重矩阵，上下文向量 c_t 是注意力权重与编码器所有隐藏层状态的加权和，即 $c_t=\sum_S\alpha_{ts}\bar{h}_s$。

模型首先经过预处理得到光流 O、行人外观表征 R_{li} 和全局上下文 C_g。光流送入时间流卷积模块得到时间流特征编码 T_O。全局上下文 C_g 与行人外观表征 R_{li} 送入 MobileNet 网络经过融合得到卷积特征编码 V_C，再经由空间注意力模块根据空间特征信息加权得到空间流特征编码 S_C。时空流特征编码分别经过 GRU 单元重编码得到时间流特征向量 F_{To} 和空间流特征向量 F_{S_C}，时空特征向量输入时间注意力模块得到关于时序特征的时间注意力向量 a_{t1} 和 a_{t2}。

最后将两个时间注意力向量拼接融合输入到模态注意力模块得到关于各模态特征加权的注意力编码 a_m，再输入 Dense 层获得行人过街动作概率 A_i^{t+n}，判断行人过街意图。模型的行人动作概率如下：

$$\begin{aligned}A_i^{t+n}&=Input\left(O,R_{li},C_g\right)\\&=f\left(T_O,S_C\right)\\&=AT\left(F_{T_O},F_{S_C}\right)\\&=sigmoid\left(a_m\left[a_{t1},a_{t2}\right]\right)\end{aligned} \tag{7}$$

其中 AT 为注意力模块，*sigmoid* 为激活函数。

4. 对比实验与结果分析

4.1. 实验设置

评估模型使用的数据集是 JAAD [17]和 PIE [18]。其中 JAAD 提供 346 个经过剪辑的短视频(5~10 秒长)，包含两个子集，分别是 jaad_beh 和 jaad_all，前者包含正在过街的行人 495 个，即将过街的行人 191 个，后者包含额外的出现在视野中远离马路或者没有动作的行人 2100 个。PIE 中包含超过 30 万张的已标注视频帧，过街行人 512 个，其他行人 1322 个。

使用标准二元分类指标评估模型结果，分别是准确率(Accuracy)，AUC、F1 分数(F1 Socre)、精确率(Precision)和召回率(Recall)。dropout=0.5，损失函数使用二元交叉熵误差，优化器使用 Adam，batchsize 为 16，学习率 4×10^{-6}。实验使用的硬件设备是显存 11G 的 NVIDIA GeForce RTX 2080 Ti，内存 16G，深度学习框架为 tensorflow。

本文模型基于 PCIP 开发，将提出的模型与同类型方法进行了比较。参与对比试验的方法包括 PCIP [9]、PCPA [8]、SF-GRU [7]、SingleRNN [19]。

4.2. 结果分析

如图 5 为模型对比试验的定性结果分析示例。t 时刻为观察到图像序列最后一帧的时间，t + 0.5 s 时刻和 t + 0.25s 时刻是 t 时刻前观察到图像序列中独立帧的时间，行人行为发生于 t + 1s 时刻，模型识别行

人意图预测 t + 1s 行人是否穿越马路，绿色为穿越，红色为不穿越。

场景(a)中行人保持静止，等待车辆通过。由于 SF-GRU 和 SingleRNN 只考虑行人动作和局部背景，认为靠近道路边缘的行人即将采取穿越行为导致其对行人意图出现误判。其余模型均能正确识别行人意图。

场景(b)中目标行人为儿童，外观表征像素面积较小，有跨入车道穿越马路的动作。SF-GRU 考虑行人动作因素比重较大，由儿童跨入道路的动作判定其将要穿越道路，导致识别结果错误。PCIP 考虑行人动作因素和全局场景上下文，由于道路中并无车辆，因此结合儿童动作判定其将要穿越道路，导致识别结果错误。其余模型均能正确识别行人意图。

场景(c)中目标行人为青少年，左侧一直有遮挡，难以判断行人数量。SingleRNN 考虑行人动作信息和局部背景，由于行人在看到车辆时动作出现停顿，导致模型的意图识别结果出现误判。PCPA 则结合行人动作、局部背景和车辆信息认为行人不会穿越道路，由于缺少全局场景上下文信息，导致其忽略了后续出现在场景中的其他行人，给出了错误的意图识别结果。其余模型均能正确识别行人意图。

可以看到三个场景中，本文模型的意图识别结果均与数据集的真实标签相同，其他模型都有识别错误的情况出现，说明本文模型的意图识别准确性优于同类型方法。

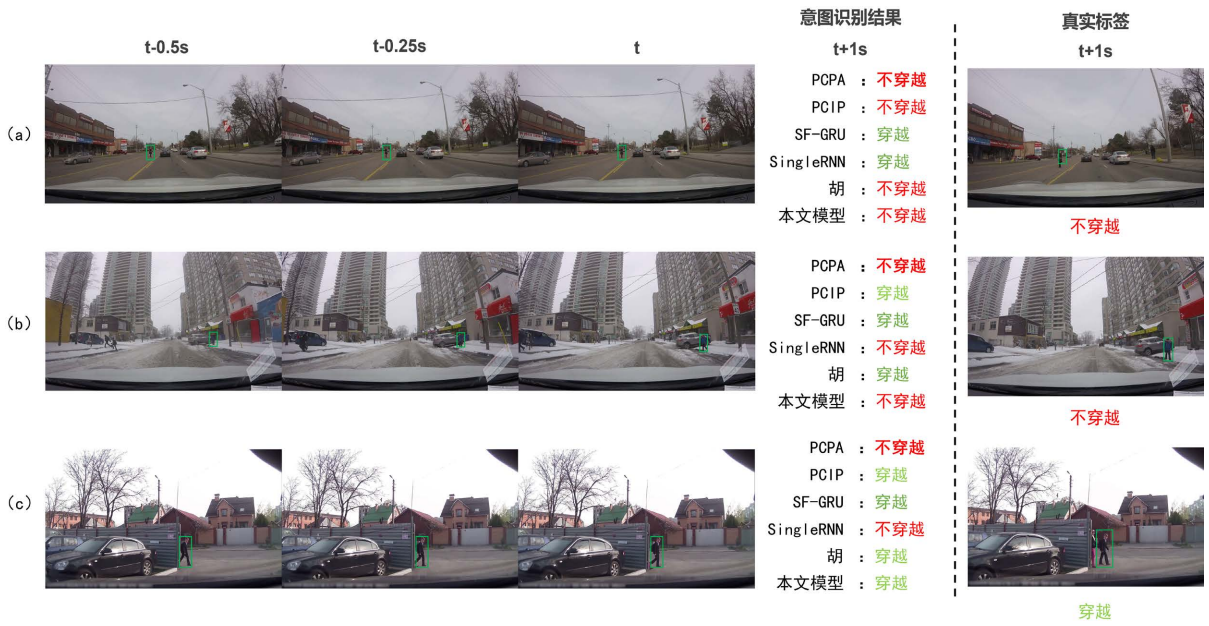


Figure 5. Comparison of qualitative analysis results
图 5. 定性分析结果对比

如表 1 所示是所有模型在数据集 jaad_beh 上的测试结果。本文方法在五项指标上都获得了最佳结果。相比对照模型中准确率最高的 SingleRNN，本文方法提高了 7 个百分点。F1 分数同时兼顾了分类模型的精确率和召回率，可以看作模型的精确率和召回率的一种加权平均，是衡量二元分类模型有效性最重要的指标。相比对照模型中 F1 分数最高的 PCIP，本文方法也提高了 4 个百分点。由于 jaad_beh 中只包含正在过街或者即将过街的行人样本，样本数量小，泛化性不足，因此在其基础上增加了出现在视野中远离马路或者没有动作的行人样本，再次进行训练和测试。

如表 2 所示是所有模型在数据集 jaad_all 上的测试结果。在 Jaad_all 中，本文提出的模型在关键的准确率、F1 分数和精确率三个指标中都优于对比模型。

Table 1. Quantitative results for the dataset jaad_beh
表 1. 数据集 jaad_beh 的定量结果

模型	模型结构		Jaad_beh				
	空间特征编码	光流	Accuracy	AUC	F1	Precision	Recall
SingleRNN	VGG + GRU	×	0.59	0.52	0.71	0.64	0.80
SF-GRU	VGG + GRU	×	0.58	0.56	0.65	0.58	0.62
PCPA	3D CNN	×	0.55	0.51	0.65	0.63	0.67
PCIP	VGG + GRU	×	0.57	0.48	0.72	0.61	0.86
OURS	MOBILE + GRU	√	**0.66**	**0.58**	**0.77**	**0.67**	**0.89**

*加粗的字体是表中对应评估指标最好的模型测试结果。

Table 2. Quantitative results for the dataset jaad_all
表 2. 数据集 jaad_all 的定量结果

模型	模型结构		Jaad_all				
	空间特征编码	光流	Accuracy	AUC	F1	Precision	Recall
SingleRNN	VGG + GRU	×	0.79	0.76	0.54	0.44	0.71
SF-GRU	VGG + GRU	×	0.76	0.77	0.53	0.40	0.79
PCPA	3D CNN	×	0.76	0.79	0.55	0.41	**0.84**
PCIP	VGG + GRU	×	0.81	**0.81**	0.59	0.47	0.81
OURS	VGG + GRU	√	**0.89**	0.80	**0.67**	**0.68**	0.67

*加粗的字体是表中对应评估指标最好的模型测试结果。

如表 3 所示是所有模型在数据集 PIE 上的测试结果。PIE 中有更广泛的行人样本，更多样的数据分布，更接近真实城市道路场景。在 PIE 中，本文提出的模型在仅在召回率一项指标上落后于对比方法 PCPA，其他指标均领先于对比方法。

Table 3. Quantitative results for the dataset PIE
表 3. 数据集 PIE 的定量结果

模型	模型结构		PIE				
	视觉编码	光流	Accuracy	AUC	F1	Precision	Recall
SingleRNN	VGG + GRU	×	0.80	0.73	0.64	0.64	0.75
SF-GRU	VGG + GRU	×	0.83	0.78	0.63	0.50	0.78
PCPA	3D CNN	×	0.85	0.86	0.77	0.69	**0.88**
PCIP	VGG + GRU	×	0.87	0.86	0.78	0.73	0.83
OURS	VGG + SSCBAM	√	**0.89**	**0.88**	**0.82**	**0.79**	0.85

*加粗的字体是表中对应评估指标最好的模型测试结果。

5. 意图识别系统实验验证

5.1. 行人过街意图识别系统设计

为了验证本文所提行人过街意图识别算法在复杂城市交通环境下的准确性、合理性和稳定性，采用

Python 联合 Tensorflow-Keras 深度学习框架设计了行人过街意图识别系统，系统主要包括自车视角图像加载模块、行人特征提取模块、特征编码模块，意图识别模块。以真实的城市道路条件为实验环境，确定硬件型号和软件参数，搭建实验平台，进行实车实验。通过意图识别系统对车辆采集到的第一人称视角的视频数据进行识别，以验证意图识别系统的稳定性和准确性。

系统使用 Python 语言进行开发，系统的人机交互界面基于 Python 3 版本中的 tkinter 模块进行编程，编程平台为 VS Code 和 Anaconda。在模型改进和设计方面，使用 Tensorflow-Keras 深度学习框架对模型进行搭建。本系统的具体软硬件开发环境如表 4 所示。

Table 4. System environment configuration
表 4. 系统环境配置

实验平台	开发环境配置	参数型号
硬件平台	CPU	InterCorei7-9750H 处理器
	GPU	NVIDIAGeForceGTX2080Ti
	内存	32 GB
软件架构	操作系统	Ubuntu 20.04 64 位
	编程平台	VisualStudio Code + Anaconda
	深度学习框架	Tensorflow-Keras
	编程语言	Python 3.8.12
	GPU 加速库	CUDNN 7.6.5
	计算平台	CUDA 10.0

使用 tkinter 模块设计行人过街意图识别系统的人机交互界面，主要包括：视频加载模块、图像预处理模块、特征卷积模块和意图识别模型加载模块。

1) 视频加载模块

该模块是系统的输入，导入车载摄像头采集的道路图像视频流。

2) 图像预处理模块

该模块负责预处理已采集的视频数据流。

3) 特征卷积模块

该模块选择行人特征对应的编码模块，包括处理光流的原始 I3D 或者 TFENet，空间特征编码模块主干网络的选择。

4) 意图识别模型加载模块

该模块选择意图识别使用的算法模型，包括第四章提到的对比模型 PCPA、PICP 和本文模型。

5.2. 实车道路实验

在上海市松江区多条开放的城市道路和校园内多条道路下进行实车实验，主要包含的道路场景为有信号灯十字路口、无信号灯十字路口、无信号灯人行横道路口和路侧行人较多的直行道路等典型交通场景。完成车载硬件布置及软件环境搭建之后，车辆以 0~40 km/h 的速度行驶在交通道路上进行行人过街意图识别实车实验。

实车实验采集的有效行人样本约为 200 个，使用所有意图识别模型对 200 个行人样本进行意图识别，并对识别结果进行了统计，统计结果如表 5 所示。

Table 5. Recognition results of different models
表 5. 不同模型的识别结果

	双流卷积多注意力	PCPA	PCIP
准确识别个数	191	179	186
准确率	96%	92%	90%
平均测试耗时	0.37 s	0.29 s	0.33 s

由表 5 可知，本文模型在 200 个新样本上的准确率领先于同类型模型，只在平均测试耗时这一项上稍微落后。说明将各个功能集成在一起之后本文模型依旧稳定，验证了本文行人过街意图识别系统的有效性。

6. 结论

正确识别城市道路场景中各种道路使用者的行为和意图对自动驾驶汽车的应用有非常积极的作用。我们提出了一种基于双流网络结构融合时空特征的行人过街意图识别模型。基于输入的视频流获取空间流和时间流，基于 MobileNet 结合空间注意力设计了空间流特征提取模块，通过优化 I3D 设计了时间流特征提取网络，搭建了基于注意力机制的双流融合网络。在 JAAD 和 PIE 数据集上的实验证明了本文方法的有效性。通过对软件功能需求分析，将各模块集成在系统中，利用 Python 语言的 tkinter 模块设计了图形用户界面，使用实车采集的 200 个新样本对系统进行了验证，结果表明本文模型意图识别准确率为 96%，领先于对比模型，验证了识别系统的稳定性与准确性。

由于光流需要提前生成，费时费力，因此未来会探索能代替光流，不需要额外工作的特征信息。行人之间的交互也是影响行人下一步动作和意图的重要因素，同样是未来的工作需要关注的重点。

基金项目

国家自然科学基金(62101314)：基于姿态超图匹配和迁移融合学习行人重识别研究。

参考文献

[1] Ahmed, S., Huda, M.N., Rajbhandari, S., *et al*. (2019) Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey. *Applied Sciences*, **9**, Article 2335. https://doi.org/10.3390/app9112335

[2] 胡远志, 蒋涛, 刘西, 等. 基于双流自适应图卷积神经网络的行人过街意图识别[J]. 汽车安全与节能学报, 2022, 13(2): 325-332.

[3] 杨彪, 范福成, 杨吉成, 等. 基于动作预测与环境条件的行人过街意图识别[J]. 汽车工程, 2021, 43(7): 1066-1076.

[4] 曹昊天, 施惠杰, 宋晓琳, 等. 基于多特征融合的行人意图以及行人轨迹预测方法研究[J]. 中国公路学报, 2022, 35(10): 308-318.

[5] Chen, T., Tian, R.R. and Ding, Z.M. (2021) Visual Reasoning Using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 3096-3102. https://doi.org/10.1109/ICCVW54120.2021.00345

[6] Lorenzo, J., Parra, I., Wirth, F., *et al*. (2020) RNN-Based Pedestrian Crossing Prediction Using Activity and Pose-Related Features. 2020 *IEEE Intelligent Vehicles Symposium* (*IV*), Las Vegas, 19 October-13 November 2020, 1801-1806. https://doi.org/10.1109/IV47402.2020.9304652

[7] Rasouli, A., Kotseruba, I. and Tsotsos, J.K. (2020) Pedestrian Action Anticipation Using Contextual Feature Fusion in Stacked RNNs. arXiv: 2005.06582.

[8] Kotseruba, I., Rasouli, A. and Tsotsos, J.K. (2021) Benchmark for Evaluating Pedestrian Action Prediction. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2021, 1257-1267. https://doi.org/10.1109/WACV48630.2021.00130

[9] Yang, D.F., Zhang, H.L., Yurtsever, E., Redmill, K.A. and Özgüner, Ü. (2022) Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention. *IEEE Transactions on Intelligent Vehicles*, **7**, 221-230. https://doi.org/10.1109/TIV.2022.3162719

[10] Cordts, M., Omran, M., Ramos, S., *et al*. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3213-3223. https://doi.org/10.1109/CVPR.2016.350

[11] Chen, L.C., Zhu, Y.K., Papandreou, G., Schroff, F. and Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv: 1706.05587. https://doi.org/10.1007/978-3-030-01234-2_49

[12] Ilg, E., Mayer, N., Saikia, T., *et al*. (2017) FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 1647-1655. https://doi.org/10.1109/CVPR.2017.179

[13] Howard, A.G., Zhu, M.L., Chen, B., *et al*. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861.

[14] Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. 15*th European Conference of Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[15] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 4724-4733. https://doi.org/10.1109/CVPR.2017.502

[16] Luong, T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. arXiv: 1508.04025. https://doi.org/10.18653/v1/D15-1166

[17] Rasouli, A., Kotseruba, I. and Tsotsos, J.K. (2017) Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. 2017 *IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 22-29 October 2017, 206-213. https://doi.org/10.1109/ICCVW.2017.33

[18] Rasouli, A., Kotseruba, I., Kunic, T. and Tsotsos, J. (2019) PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6261-6270. https://doi.org/10.1109/ICCV.2019.00636

[19] Kotseruba, I., Rasouli, A. and Tsotsos, J.K. (2020) Do They Want to Cross? Understanding Pedestrian Intention for Behavior Prediction. 2020 *IEEE Intelligent Vehicles Symposium* (*IV*), Las Vegas, 19 October-13 November2020, 1688-1693. https://doi.org/10.1109/IV47402.2020.9304591